Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(1), 2014, 80-86

ISSN 2347 – 7393

Original Research Article

SURVEY ON ASSOCIATION RULE MINING USING "APRIORI" ALGORITHM

Pramod B Warale*[1], Yogesh S Khaladkar* [2]

[1]TE Comp PCCOE
[2]TE DYPIET

Abstract- Association rule mining is the important technique in the field of data mining. The important task of association rule is to find the frequent item set .In frequent pattern mining, there are several algorithms. Apriori is classical and most famous algorithm .Objective of using Apriori algorithm is to find frequent item set and association between different item set that is association rule .It is simple algorithm but, it has some drawback in case of large database, if database is large, it required more scan. So it required more space and time. There are so many researchers have been done for improvement of this algorithm. So this paper presents a survey on association rule mining using Apriori algorithm, of recent research work carried by different researchers

Keywords: Association rule, Data mining, Apriori algorithm, Support, Confidence.

1. Introduction

Population increases day by day as the population increases day by day database also increasing an alarming rate. As well as Information Technology is growing, databases created by the organizations are becoming huge. The organization sectors include banking, manufacturing, transportation, marketing, telecommunications etc. We need to extract valuable data completely and efficiently. There is complex query which are difficult to retrieve large database so we need some technique such as association rule mining algorithm using Apriori algorithm.

I**1.1 Data mining:-** The process of extracting valid, previously unknown, compressible, and actionable information from large databases and using make valuable business decision[8]. KDD (Knowledge Discovery in Database) and data mining are used as synonyms to each other. But in real, Data mining is core process of KDD.

1.2 Association rule:-Association rule are used in finding relationship between any database Association rule are derived from finding all frequent item set by using the minimum support and using confidence value, also strength of association rule is measured in term of support and confidence. Association rule is the implication expression in the form of

A→B, where A&B is disjoint item set.

A. Support (S):- Support are used to determine the how item frequently occurs in database. Support determines the how often rule is applicable to given dataset. Its support is the percentage of transaction in database that contains AUB (means A and B) [1].Support can be calculated by using the formula given below Support (AB) = Support count of (A∪B)/Total number of transactions in database.

B. Confidence (C):-In the confidence we find the probability that items are occurs together in the transaction .In confidence determine how frequently items in B appear in instance that contain A .It gives the percentage of transaction in database containing A that also contain B[1]. Confidence can be calculated by using formula, Conf (AB) =Support count of (A∪B) / Support (A) **C. Lift:-**The lift of rule is defined as: Lift (A→B) = Supp (AUB)/Supp (B)*Supp (A) [1].

1.3 Depend on number of data dimensions involved in the rule

1.3.1 Single dimensional association rule:-if item or attributes refer only one dimension [7].

1.3.2 Multidimensional association rule:-if item or attributes refer two or more dimensions [7].

Apriori Algorithm

"Apriori=prior knowledge". Apriori is the Latin word and it meaning is "From what come before". Apriori uses bottom up strategy [1].It is the most famous and classical algorithm for mining frequent patterns

Frequent Item set:-These are the item set that satisfy minimum support threshold. Apriori property is any set of frequent item set must be frequent. Apriori makes use of an iterative approach known as breath-first search, where k-1 item set are used to search k item sets. There are two main steps in Apriori execution as shown in fig 1.

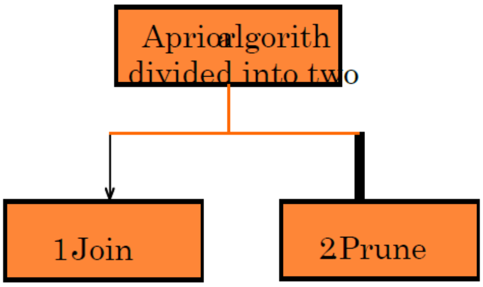


Figure1. Apriori in two steps

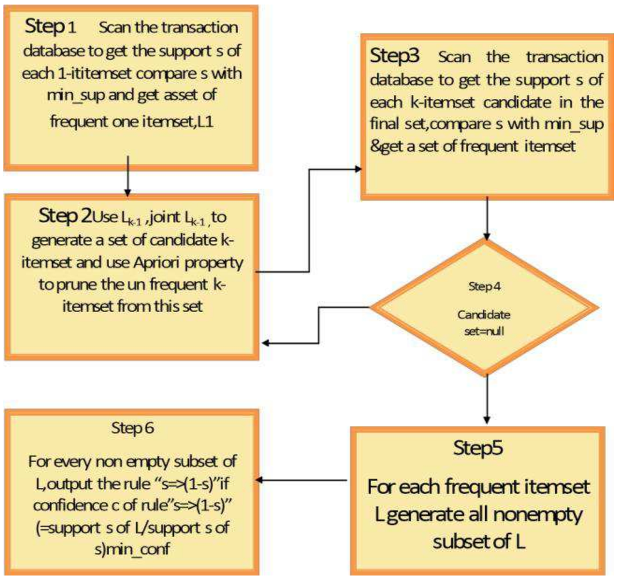
2.1 The Apriori Algorithm: Pseudo code [5]

•Join Step: Ck is generated by joining Lk-1with itself

•Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent kitem set

•Pseudo-code:

```
Ck: Candidate item set of size k Lk: frequent item set of size k
L1= {frequent items};
For (k=1; Lk! = 1; k++) Do Begin
Ck+1= candidates generated from Lk;
For each transaction tin database do
Increment the count of all candidates in Ck+1that are contained in t
Lk+1= candidates in Ck+1with min_support
End
Return k∪Lk
```

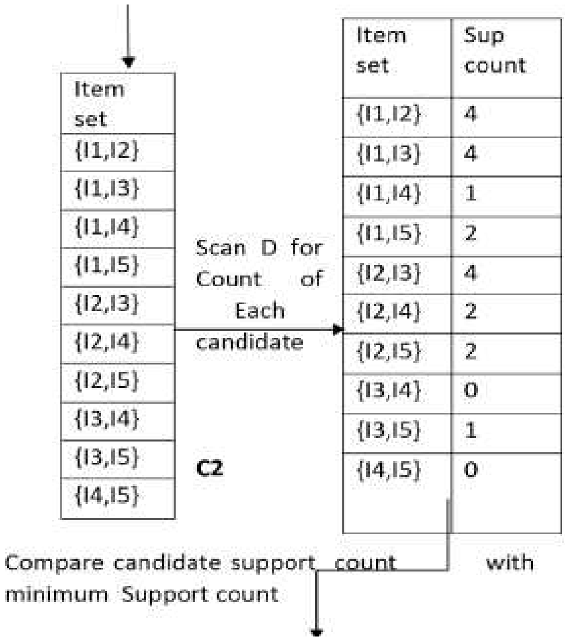

Step 1 Scan the transaction database to get the support s of each 1-ititemset compare s with min_sup and get asset of frequent one itemset,L1
Step 2 Use L_{k-1} ,joint L_{k-1} ,to generate a set of candidate k-itemset and use Apriori property to prune the un frequent k-itemset from this set
Step3 Scan the transaction database to get the support s of each k-itemset candidate in the final set,compare s with min_sup &get a set of frequent itemset
Step4
Candidate set=null
Step5
For each frequent itemset L generate all nonempty subset of L
Step 6
For every non empty subset of L,output the rule "s=>(1-s)"if confidence c of rule"s=>(1-s)" (=support s of L/support s of s)min_conf

2.2 Flowchart:-the flowchart gives overall idea About Apriori .as shown in figure 2.

Example

Consider a database, consisting of 9 transactions. Suppose min support count required is 2(i.e min_sup=2/9=22%).Let minimum confidence required is 70%.We have to first find out the frequent item set using Apriori algorithm. Then, Association rules will be generated using min. support & min.

TID	List of Item
T100	I1,I2,I5
T100	I2,I4
T100	I2,I3
T100	I1,I2,I4
T100	I1,I3
T100	I2,I3
T100	I1,I3
T100	I1,I2,I3,I5
T100	I1,I2,I3



Item set
{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}
{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}

Item set	Sup count
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

Scan D for Count of Each candidate
C2
Compare candidate support count with minimum Support count

Step1:-Generating 1- Item set frequent pattern.

The set of frequent 1-itemsets, L1, consists of the candidate 1-itemsets satisfying minimum support. In the first iteration of the algorithm, each item is a member of the set of candidate.

Scan the D for count of each candidate

Item set	Sup count
I1	6
I2	7
I3	6
I4	2
I5	2

C1

Compare candidate Support Count with Minimum support Count.

Item set	Sup count
I1	6
I2	7
I3	6
I4	2
I5	2

Item set	Sup count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Step 2:- Itemsets discover itemset, uses L1, candidates C2 Next, are support count for each candidates item is C2 is shown in the middle table). The set of frequency 2-determined, consisting of those candidates 2L2.

Generating 2- frequent pattern :- to the set of frequency 2-L2, the algorithm ojin L1 to generate a set of 2- itemsets, scanned and the accumulated (as itemsets, L2, is then determined consisting of those candidates 2L2.

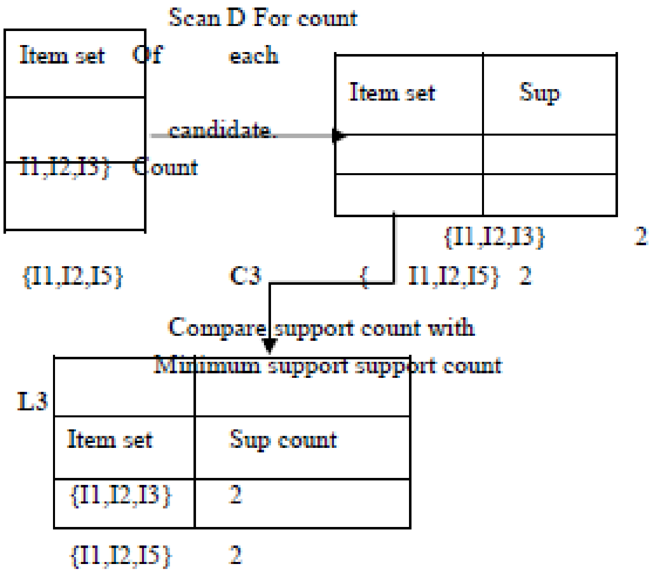
Step3:-Generating 3_Itemset frequent pattern

The generation of the set of candidate 3 item sets, C3, involves use of the Apriori Property. In order to find C3, we compute L2*Join*L2.C3= L2

*Join*L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}. Now, Join steps complete and Prune step will be used to reduce the size of C3.

Prune step helps to avoid heavy computation due to large Ck. Based on the Apriori property that all subsets of a frequent item set must also be frequent; we can determine that four latter candidates cannot possibly be frequent. How?

- For example, let's take {I1, I2, I3}.The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, I3} are members of BUT, {I3, I5} is not a member of L2and hence it is not frequent violating Apriori Property.
- Therefore, C3= {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning.
- Now, the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3having minimum support confidence.



Step 4:- Generating 4_Itemset frequent pattern.

The algorithm uses L3*Join*L3 to generate a candidate set of 4-itemsets, C4. Although the join results in {{I1, I2, I3, I5}}, this item set is pruned since its subset {{I2, I3, I5}}is not frequent. Thus, C4= φ, and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm. These frequent item sets will be used to generate strong association rules (where strong association rules satisfy both minimum support & minimum confidence).

Step 5: Generating Association Rules from Frequent Item sets

•Procedure:

•For each frequent item set ***"l"***, generate all nonempty subsets of ***l.***

•For every nonempty subset ***s*** of ***l***, find the min

•Back to Example:

We had L={{I1},{I2},{I3,{I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.

Let's take ***l*** = {I1,I2,I5}.

It's all nonempty subsets are {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}.

Let minimum confidence thresholds', say 70%.

The resulting association rules are shown below, each listed with its confidence.

R1: I1 ^ I2→I5

• Confidence = sc {I1,I2,I5}/sc{I1,I2} = 2/4 = 50%

•R1 is rejected.

R2: I1 ^ I5 →I2

•Confidence = sc {I1,I2,I5}/sc{I1,I5} = 2/2 = 100% •R2 is selected.

R3: I2 ^ I5 →I1

•Confidence = sc {I1,I2,I5}/sc{I2,I5} = 2/2 = 100% •R3 is selected
R4: I1 →I2 ^ I5
•Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%
•R4 is rejected.
R5: I2 →I1 ^ I5
•Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29%
•R5 is rejected.
R6: I5 →I1 ^ I2
•Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%
•R6 is selected.
In this way, we have found three strong association rules.

Survey

Mining Efficient Association Rules through Apriori Algorithm Using Attribute and comparative Analysis of Various Associations Rule Algorithms [1]

Author consider data as bank data .They have obtained result for three different association rule that is Apriori association rule, Predictive Apriori association rule & Tertius association rule. By using Weka a data mining tool .According to the result obtained using data mining tool author find that Apriori association algorithm perform better than Predictive Apriori association rule and Tertius association rule algorithm .some of the disadvantages are there that is In case of such large database Apriori is not efficient algorithm because 1)It requires multiple scan over the database to generate candidate set.2)It also takes more memory ,space and time.3)Algorithm scan database repeatedly for searching frequent item set ,so more time and resource are required in large number of scans. By using the above three algorithm we can overcome the some problem of Apriori algorithm.

An Enhanced Scaling Apriori Algorithm to minimize the number of candidate sets while Generating Association Rule [2] This paper used for the following two purposes.

The quantitative association rule mining with the enhancement on Apriori algorithm. The algorithm for generating quantitative association rules start by counting the item ranges in the databases, in order to determine the frequent ones. These frequent item ranges are the basis for generating order item ranges using an algorithm similar to Apriori, taking into account the size of transaction as the number of items that it comprises. **b.** The reduction of memory utilization during the pruning phase of the transactional execution. This part of the algorithm generate all candidates based on 2_frequent set on sorted database, and all frequent item set that can no longer be supported by transaction that still have to be processed

Thus the new algorithm no longer has to maintain the covers of all past item sets into main memory .In this way, proposed levelwise algorithm accesses a database less often than Apriori and requires less memory because of utilization of additional upward closure properties.

The Research of Improved Association Rules Mining Apriori Algorithm [4]

To overcome on the time required for database Scan author presents new algorithm that is improved association rule mining. In this algorithm he uses hash tree to store the candidate item set. Depending on the result it found that the time required for processing is decreased.

An Improved Apriori Algorithm for Association Rules of Mining [5]]

In this paper author proposed new algorithm firstly while scanning the database it separate every acquired data depending on the discretization of data items and count of data.

Then prune the acquired item set. After the analysis it reduces system resources ocuupied.

Conclusion

According to the result obtained using data mining tool author find that Apriori Association algorithm performs better than the Predictive Apriori Association Rule and Tertius Association Rule Algorithms [1]. By using Apriori algorithm, association rule are very useful in application of banking & market basket analysis. It presents a remarkable advantage in reducing the feature number, due to using Apriori algorithm directly to mining association rule [3]

Future scope

From survey it is conclude that in the Apriori algorithm as the databases increases the no of scan increases so there are many improvements are needed on time and space to make Apriori efficient. Improve the time and space is the area to work on. Association rule produced by Predictive Apriori algorithm, Tertius algorithm and Apriori association rule algorithm can be combined for better result.

References

1. MS Shweta, Dr .Kanwal Garg, "Mining Efficient Association Rules through Apriori Algorithm Using Attribute and comparative Analysis of Various Association Rule Algorithm "In: volume 3, Issue 6, June 2013 ISSN: 2277 128X
2. International Journal of Advanced Research in Computer science and software Engineering.